Solution

25 June 2021 20:02

Solution to PCA

$$X = \begin{array}{c|c} & 1, 2, \ldots, d \\ \begin{array}{c} 1 \\ 2 \\ \vdots \\ n \end{array} & \left[\quad \right]_{n \times d} \end{array} \qquad S = \mathrm{cov}(X) = \begin{array}{c|c} & 1, 2, \ldots, d \\ \begin{array}{c} 1 \\ 2 \\ \vdots \\ d \end{array} & \left[\quad \right]_{d \times d} \end{array} = \boxed{\frac{1}{(n-1)}(X^T X)}$$

(Column standardized)

Because mean = 0

eigen values of $\boxed{\mathrm{cov}(X) = S}$ are $\rightarrow \lambda_1, \lambda_2, \lambda_3, \ldots \ldots \lambda_d$

Corresponding eigen vectors $\rightarrow V_1, V_2, V_3, \ldots \ldots V_d$

$$\boxed{\underbrace{S \; V_1}_{d\times 1} = \underbrace{\lambda_1 V_1}_{d \times 1}}$$

S: $d \times d$, V_1: $d \times 1$, V_1: $d \times 1$

$V_i \perp^r V_j \Rightarrow V_i^T V_j = 0$

$u_j = V_j$ = eigen Vector of S corresponding to eigen value λ_j

Summary

$X = [\quad]$

1. Column Standardization of X
2. $\mathrm{cov}(X) = S = \frac{1}{(n-1)} X^T X$
3. Calculate eigen values & eigen vectors of S.

   $\lambda_1 \geq \lambda_2 \geq \lambda_3 \geq \lambda_4 \geq \ldots \ldots \geq \lambda_d$

   $V_1 \quad V_2 \quad V_3 \quad V_4 \quad \ldots \ldots \quad V_d$

4. $u_1 = V_1$, $u_2 = V_2$, ...

   choose first d' eigen vectors to project d-dim data to d' - data space. ($d' \lll d$)

   Just dot product of eigen vector with X, will transform d-dim to d'-dim

Interpretation of eigen values:—

... will tell about explained variance:

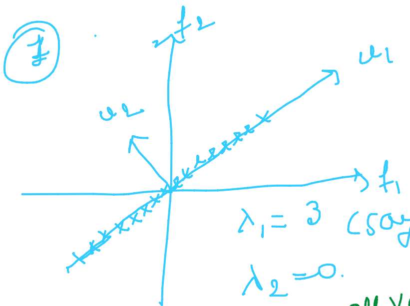
Interpretation of eigen values :—

Eigen values will tell about explained variance :

$$\boxed{\frac{\lambda_i}{\sum_{j=1}^{d} \lambda_j}}$$

this will give % of variance explained/preserved by eigen vector corresponding to λ_i .

(1)



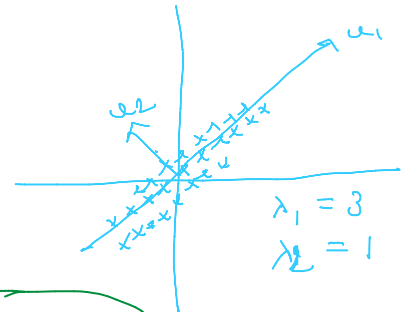
$\lambda_1 = 3$ (say)

$\lambda_2 = 0$.

all variance preserved along u_1

Here, $\frac{\lambda_1}{\lambda_1+\lambda_2} = \frac{3}{3} = 100\%$

(2)



$\lambda_1 = 3$

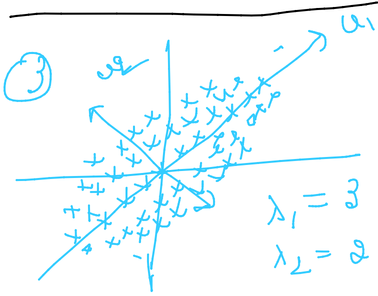
$\lambda_2 = 1$

$$\boxed{\frac{\lambda_1}{\lambda_1+\lambda_2} = \frac{3}{4}}$$

75% variance preserved along u_1 & remaining along u_2

---

(3)



$\lambda_1 = 3$

$\lambda_2 = 2$

Here, $\frac{\lambda_1}{\lambda_1+\lambda_2} = \frac{3}{5} = 60\%$

60% of variance explained by u_1

(4)

$\lambda_1 = 3$

$\lambda_2 = 3$

Here

$\frac{\lambda_1}{\lambda_1+\lambda_2} = \frac{\lambda_2}{\lambda_1+\lambda_2} = \frac{1}{2} = 50\%$

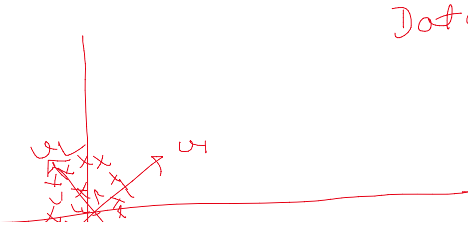
u_1 preserves 50% of variability

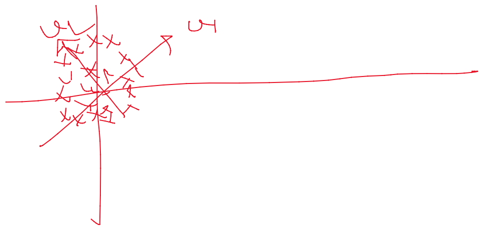
u_2 also preserve 50% of variability

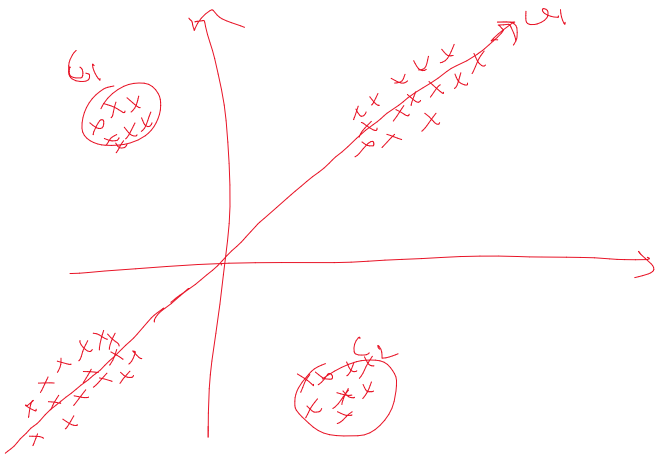
LIMITATIONS :—

Data is spherically distributed then $\lambda_i \sim \lambda_j$,

So More info will be lost if we do PCA .

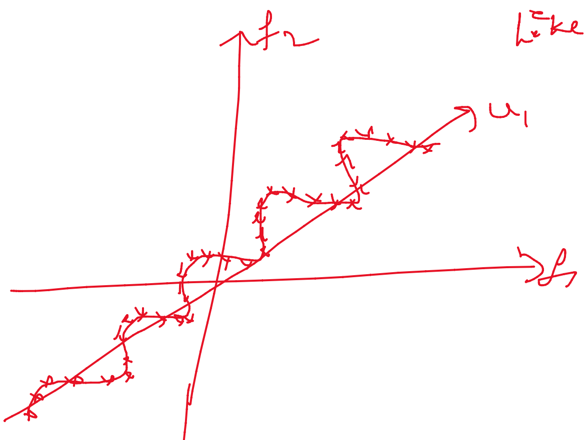






Here when we perform PCA,
c_1 & c_2 will be projected at origin
Then we can't distinguish c_1 & c_2.

Here PCA <u>fails</u>

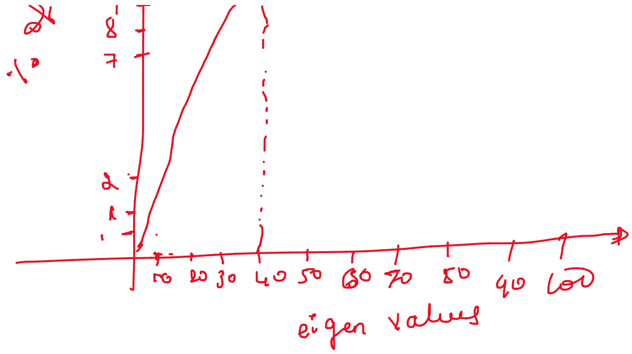


Like sine wave.
Here we can find u_1, but
we can preserve the shape of sine wave.

These are all limitations of PCA

We use PCA to reduce dimension from d to d' $(d' <<< d)$
while reducing based on required % of preservation of variance,
we can choose value of d'.

⇒ Here mostly 95% Variance is explained
by 40 features



by 40 features
Rest 60 feature contributing only
5% variance.

Hence we can reduce d (100) to
d' (40).

This plot will help so much in real world.

This is all about PCA

Scikit learn has direct function to perform PCA.

We can implement PCA without sklearn.
Simple code illustration is attached

REVIEWED
By Praveen Hegde at 7:03 pm, Aug 29, 2021